

17.1 A 0.6V 1.5GHz 84Mb SRAM Design in 14nm FinFET CMOS Technology

Eric Karl, Zheng Guo, James W. Conary, Jeffrey L. Miller, Yong-Gee Ng, Satyanand Nalam, Daeyeon Kim, John Keane, Uddalak Bhattacharya, Kevin Zhang

Intel, Hillsboro, OR

The growth of battery-powered mobile and wearable devices has increased the importance of low-power operation and cost in system-on-a-chip (SoC) design. Supply-voltage scaling is the predominant approach to active power reduction for SoC design, including voltage scaling for on-die memory given increasing levels of memory integration. SRAM can limit the minimum operating voltage (V_{\min}) of a design, often leading to the introduction of separate voltage supplies for on-die memory. Additional supplies increase platform cost, and operating memory at higher voltage leads to increased power consumption. The introduction of trigate devices at the 22nm technology node delivered superior short channel effects and subthreshold slope relative to existing bulk planar device technology enabling reduction in threshold voltage within a fixed leakage constraint. Lower transistor V_{th} , improvements to random device variability, and assist circuits to overcome device-size quantization enabled a >150mV reduction in SRAM V_{\min} [1]. At the 14nm technology node, FinFET device-size quantization remains a challenge for compact 6T SRAM bitcells with minimum-size transistors. Careful co-optimization between technology and design of memory-assist circuits is required in order to deliver dense, low-power memory operation at low voltages. In this paper, we present an 84Mb SRAM array design with wide-voltage-range operation in a 14nm logic technology featuring 2nd-generation FinFET transistors.

Figure 17.1.1 shows a layout diagram of a 0.0500 μm^2 high-density 6T SRAM cell (HDC) and a 0.0588 μm^2 low voltage 6T SRAM cell (LVC) in 14nm FinFET technology [2]. HDC has a fin ratio of 1:1:1 (PU:PG:PD) and LVC features a larger PD device at 1:1:2, to provide improved read stability and performance at low voltage. Self-aligned double-patterning techniques have extended the capabilities of 193nm immersion lithography on critical layers to deliver 0.54 \times scaling relative to comparable 22nm SRAM bitcells. Contacted gate pitch is 70nm and fin pitch is 42nm for the technology. Figure 17.1.1 highlights recently reported SRAM designs from 16nm, 14nm and 10nm technologies [1-6]. The 0.050 μm^2 HDC cell in this work is the smallest reported SRAM cell at any technology node among the cited work. Despite significant geometric scaling from the 22nm node, optimizations to fin profile and subfin doping on the 2nd-generation FinFETs enable a nearly 2 \times reduction in device random threshold voltage variation [2], a critical factor for 6T SRAM V_{\min} .

The HDC SRAM cell has a 1:1 PG to PD ratio, leading to degraded stability from charge injection during read operations. Wordline underdrive (WLUD) is utilized in the HDC arrays as an area-efficient approach to enhance read-stability margin at the cost of cell performance [1]. Suppressed BL techniques, such as the DNR circuit, also improve read stability, but are effective across a limited range of process technology targets and the implementation leads to more area overhead and higher power consumption than WLUD [3]. The LVC SRAM, with a 1:2 PG to PD ratio, has adequate read-stability margin and does not require WLUD, enabling higher performance at low voltage. Both HDC and LVC have a 1:1 ratio between the PG and PU devices, which limits write margin at low voltage due to contention between these two devices, motivating the use of a write-assist circuit. In Figure 17.1.2, a column-based TVC circuit is shown, utilizing an NMOS device to discharge the memory cell supply (VCS) to weaken the PU transistor during writes [1]. Half-selected cells along the written column face an instability risk if VCS is pulled below the minimum retention voltage. This can be mitigated by controlling the amplitude of the supply-voltage collapse in this circuit by enabling devices connected to VCC that clamp VCS to an intermediate level. The downside in this implementation is static current flow between supplies when clamping VCS. In this 14nm FinFET technology, improvements to random variation reduce the amplitude of VCS collapse required to achieve low- V_{\min} operation. The conventional TVC circuit topology requires the highest static current to clamp VCS near VCC, limiting the realizable write-energy reduction.

Figure 17.1.3 details a charge-share transient voltage collapse circuit (CS-TVC) that delivers high performance at low voltage, compatibility with WLUD across a wide range of technology targets and reduces active power consumption relative

to the TVC circuit in Fig. 17.1.2. The CS-TVC switch circuits are located at the edges of a 256-row column, breaking the memory cell supply into 2 distinct 128b regions per column. Multiple switch circuits within a column are connected to a CS-TVC capacitor to reduce area, as only one switch within the column is active during write. The CS-TVC capacitor contains a primary node and 2 secondary nodes enabled by CSCAP[1:0] to modulate the effective capacitance. NMOS devices driven by the DISCHARGE signal are used to reset the GCSCAP node to VSS before a write operation. The CS-TVC operation begins with a self-timed pulse (TVC PULSE) aligned to the rising edge of WL that simultaneously disconnects the selected VCS region from VCC, disables the NMOS devices in the CS-TVC capacitor, and connects VCS to the pre-discharged GCSCAP node. Charge is balanced through a PMOS switch between the VCS region selected and GCSCAP, resulting in a temporary suppression of the VCS node to improve write margin. The falling edge of the TVC pulse completes the operation and restores the VCS voltage level to VCC.

Figure 17.1.4 shows the organization of the 68kB LVC block and 17kB LVC subarray, featuring 258b/BL and 136b/WL in a butterfly array configuration. In the logical I/O, 4 CS-TVC switch regions share one CS-TVC capacitor bank positioned in the center of the column. The 17kB LVC sub-arrays with conventional TVC and CS-TVC achieve bit densities of 11.6 and 11.3 Mb/mm² respectively. The array efficiency is 71.6% and 69.8%, respectively for conventional TVC and CS-TVC, with the CS-TVC circuitry adding 6.5% to the 17kB sub-array area. The HDC array features a denser 512b/BL column I/O design with a bit density of 14.5 Mb/mm² and 76.2% array efficiency.

In Figure 17.1.5, 50MHz and 1GHz LVC SRAM write- V_{\min} measurements are shown for different TVC methods. CS-TVC features a charge-share collapse to 48% of VCC and conventional TVC with a strong bias setting (SB-TVC) provides a clamped voltage level at 44% of VCC using the circuit from Fig. 17.1.2. Pulsed TVC (P-TVC) features a shallow collapse defined by a narrow pulsewidth, achievable by either the CS-TVC or the conventional TVC circuit. The deeper collapse of CS-TVC and SB-TVC enable 40mV lower V_{\min} at 50MHz. At 1GHz, the deeper collapse and required recovery for these designs increases V_{\min} by 30mV for CS-TVC and 75mV for SB-TVC relative to P-TVC. By eliminating static crowbar current, the CS-TVC circuit reduces active energy by 24% relative to SB-TVC for a comparable V_{\min} and VCS collapse level. P-TVC has write-energy overhead of 43% relative to no assist, but is difficult to implement across a range of array configurations.

Figure 17.1.6 shows the 14nm LVC V_{\min} is 0.6V at the 90th percentile, an 80mV reduction compared with a comparable LVC array on 22nm [1]. The 14nm HDC V_{\min} is 0.7V at the 90th percentile, within 15 to 20mV of the 22nm LVC SRAM V_{\min} . The voltage-frequency shmoo of LVC SRAM at 95°C demonstrates 1.5GHz performance at 0.6V and wide-range operation to 1V. Figure 17.1.7 shows a die micrograph of a 14nm test vehicle with HDC and LVC SRAM arrays with fuse, PLL and PBIST circuitry.

References:

- [1] E. Karl *et al.*, "A 4.6GHz 162Mb SRAM Design in 22nm Tri-Gate CMOS Technology with Integrated Active V_{\min} -Enhancing Assist Circuitry", *ISSCC Dig. Tech. Papers*, pp. 230-231, Feb. 2012.
- [2] S. Natarajan *et al.*, "A 14nm Logic Technology Featuring 2nd-Generation FinFET Transistors, Air-Gapped Interconnects, Self-Aligned Double Patterning and a 0.0588 μm^2 SRAM cell size", *IEDM Dig. Tech. Papers*, in press, 2014.
- [3] T. Song *et al.*, "A 14nm FinFET 128Mb 6T SRAM with V_{\min} -Enhancement Techniques for Low-Power Applications", *ISSCC Dig. Tech. Papers*, pp. 232-233, Feb. 2014.
- [4] Y.-H. Chen *et al.*, "A 16nm 128Mb SRAM in High-K Metal-Gate FinFET Technology with Write-Assist Circuitry for Low- V_{\min} Applications", *ISSCC Dig. Tech. Papers*, pp. 238-239, Feb. 2014.
- [5] H. Pilo *et al.*, "A 64Mb SRAM in 22nm SOI Technology Featuring Fine-Granularity Power Gating and Low-Energy Power-Supply Partition Techniques for 37% Leakage Reduction", *ISSCC Dig. Tech. Papers*, pp. 322-323, Feb. 2013.
- [6] K.-I. Seo *et al.*, "A 10nm Platform Technology for Low Power and High Performance Application Featuring FINFET Devices with Multi Workfunction Gate Stack on Bulk and SOI", *VLSI Tech. Dig. Tech. Papers*, pp. 12-13, 2014.

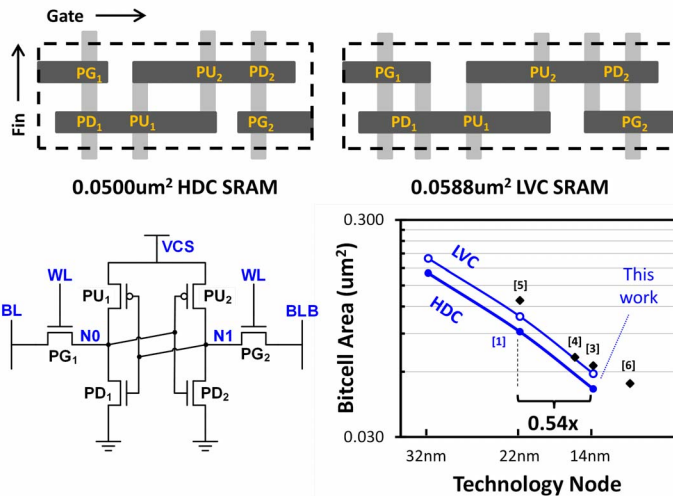


Figure 17.1.1: 14nm HDC and LVC SRAM bitcells.

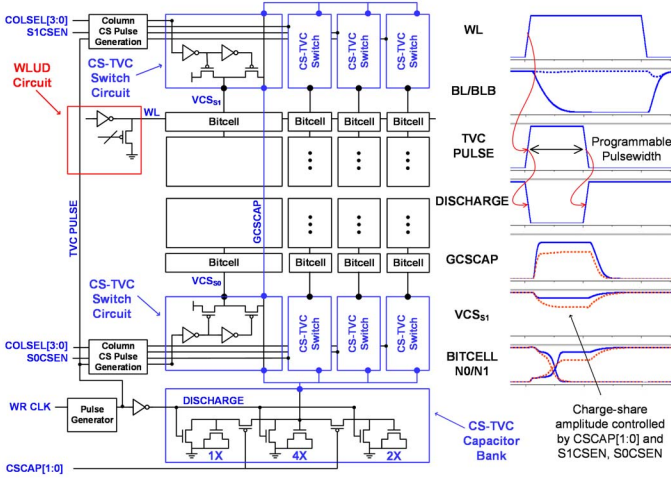


Figure 17.1.3: Charge-share transient voltage collapse circuit (CS-TVC) and operation waveforms.

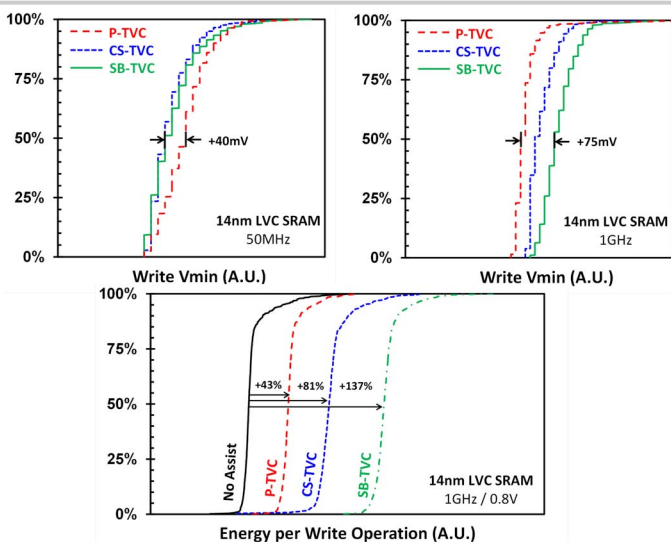
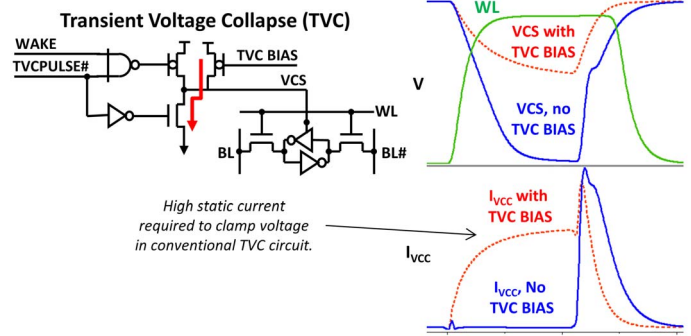
Figure 17.1.5: P-TVC, SB-TVC and CS-TVC V_{MIN} and write energy comparison.

Figure 17.1.2: Conventional transient voltage collapse (TVC) write assist circuit.

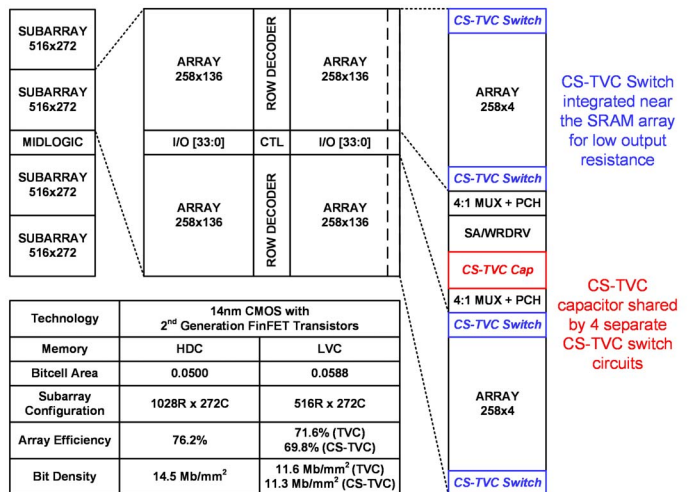
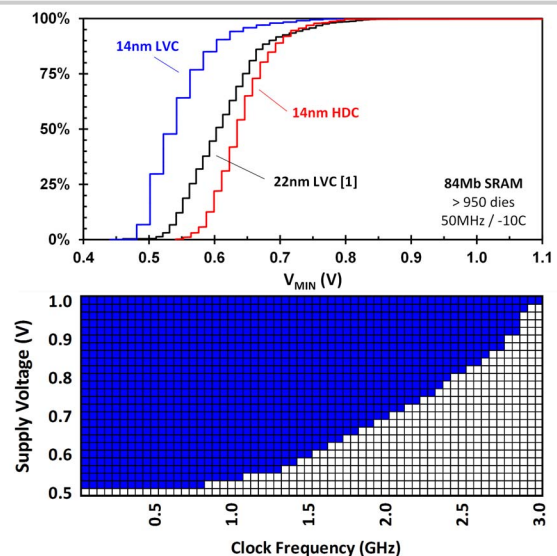


Figure 17.1.4: 548Kb SRAM architecture and density summary.

Figure 17.1.6: HDC and LVC V_{MIN} and LVC voltage-frequency shmoos.

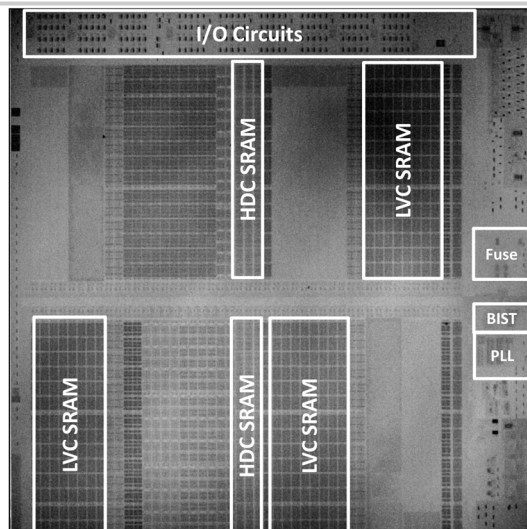


Figure 17.1.7: Die micrograph of 14nm test chip with HDC and LVC SRAM arrays.